Incremental Document Clustering for Web Page Classification

Wai-chiu Wong and Ada Wai-chee Fu
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
{wcwong1, adafu}@cse.cuhk.edu.hk

July 1, 2000

Abstract

Motivated by the benefits in organizing the documents in Web search engines, we consider the problem of automatic Web page classification. We employ the clustering techniques. Each document is represented by a feature vector. By analyzing the clusters formed by these vectors, we can assign the documents within the same cluster to the same class automatically. Our contributions are the following: (1) We propose a feature extraction mechanism which is more suitable to Web page classification. (2) We introduce a tree structure called the DC-tree to make the clustering process incremental and less sensitive to the document insertion order. (3) We show with experiments on a set of Internet documents from *Yahoo!* that the proposed clustering algorithm can classify Web pages effectively.

Keywords: Incremental update, Tree, Document, Clustering, Web, Classification

1 Introduction

The popularity of the Internet has caused a continuous massive increase in the amount of Web pages (or Web documents) [1]. The information explosion has led to a growing challenge for information retrieval systems, such as the Internet search engine, to efficiently and effectively retrieve this information. Document classification becomes an important process for helping the information retrieval systems organize this vast amount of data. For instance, most of the Internet search engines will divide the indexed Web documents into a number of classes for the users to limit the search scope. Moreover, document classification can make the retrieved results easier to browse. Most search engines return a large amount of results when a simple query (one or two words) is submitted. It can be difficult for the users to find the documents they want. If the results returned by the Internet search engine have been classified into several groups with some descriptions, the users can choose the interesting group(s) to continue browsing.

Traditionally, the document classification task is carried out manually. In order to assign a document to an appropriate class, people would analyze the contents of the document first. Therefore a large amount of human effort would be required. There has been some research work conducted on automatic text classification. One approach is to learn the text classifiers by using the machine learning techniques [9, 12, 18, 21]. However, these algorithms are based on a set of positive and negative training examples for learning the text classifiers. The quality of the resulting classifiers highly depends on the fitness of the training examples. There are many terms and classes in the World Wide Web (or just the Web), and many new terms and concepts are created everyday. It is quite impossible to have domain experts to identify training examples to learn a classifier for each text class in the above manner.

In order to make the document classification process automatic, clustering techniques have been employed. The attractiveness of cluster analysis is that it can find clusters directly from the given data, without relying on any pre-determined information such as training examples provided by domain experts. A document is commonly represented by a feature vector. Typically, each feature corresponds to a key word or phrase appearing in the set of documents. Each entry of the vector stores a numeric weight for the corresponding feature of the document. After extracting the feature vectors of the documents, we can apply the clustering algorithm on this set of the vectors as in conventional high dimensional data clustering. The resulting document clusters together with the representative features (i.e., the key words or phrases with enough *document support* within the cluster) are thus reported to the user.

Many clustering algorithms have been introduced for high dimensional data mining, for example, K-Means, CLARANS [14], BIRCH [24], CURE [6], CLIQUE [1] and ENCLUS [4]. However, the existing clustering algorithms in general are not suitable to solve the document clustering problem. Most of them (such as K-Means, CLARANS and BIRCH) require the supply of the number of clusters to work on. Unfortunately, the number of clusters or classes of the document set is usually unknown to user. Also, a major issue that document databases are now facing is the high rate of update. Researchers have found that existing clustering algorithms are not suitable for maintaining clusters in such a dynamic environment, and they have identified the problem of updating clusters without frequently performing complete re-clustering [2, 3]. Here, we propose to make use of a tree structure for incremental document

[1] We shall use the terms *Web pages* and *Web documents* and *documents* interchangeably.

clustering.

In this paper, we introduce a tree structure called the **DC-tree** (Document-Clustering Tree) which can cluster documents without any training set. With the DC-tree, an incoming data object is not forced to be inserted into the lower level when there does not exist a child node that is similar enough to this data object. This prevents some dissimilar data from being put together. As a result, the DC-tree-based clustering algorithm is less sensitive to the document insertion order and is more tolerant towards noise documents.

Feature selection is a major problem related to document clustering. We find that traditional feature extraction methods are not suitable in the Web domain. These feature extraction methods select the n highest-weighted terms as the features. The term weighting scheme is commonly based on the term frequency (TF) or the term frequency combined with the inverse document frequency (TF-IDF). However, we discover from experiments (see Section 3.1) that a Web page usually contains a small number of words and most words appear only one to two times. Hence the term frequency of a word cannot indicate the actual importance of this word. This method also ignores the *coverage* [2] of the features extracted. The coverage is important because we do not want many tiny clusters to appear. We propose a feature extraction algorithm which is more suitable for the problem domain.

We believe that our method can be useful in a number of ways: (1) For preprocessing on Web page classification so that users can choose a suitable class before searching, this helps the search to be more focused and efficient. (2) For online classification, so that when a large number of results are returned from a search, the technique can classify the results and provide better guidance to user for further searching. (3) For incremental Web page classifications after updating on the repository.

This paper is organized as follows. Section 2 is an assessment of previous work on the document clustering according to the requirements of Web document classification. Section 3 gives the background information and introduces our feature extraction method. In Section 4, the DC-tree structure and corresponding insertion algorithm will be described in detail. Section 5 reports on the experimental evaluation of our proposed feature extracting method and the Web document clustering algorithm. We conclude by summarizing our contributions in Section 6.

2 Related Work

Clustering is a useful technique in data mining for discovering interesting data distributions and patterns in the underlying data. The main advantage of using this technique is that interesting structures or clusters can be found directly from the data without using any background knowledge [8]. Current clustering techniques can be broadly classified into two categories: partitional and hierarchical. The popular K-Means and K-Medoids methods are partitional methods which maintain k cluster representatives and assign each object to the cluster with its representative closest to the object. CLARANS [14], and BIRCH [24] can be viewed as extensions of this approach to work against large databases.

BIRCH is a dynamical and incremental method to cluster the incoming points. An im-

[2] The *coverage* is the percentage of documents that contain the extracted features.

portant idea of BIRCH is to summarize a cluster of points into a clustering feature vector. This summary uses much less storage than storing all data points in the cluster. A CF-tree is built which splits dynamically. Clusters are stored in the leaf nodes. In this way, the memory requirement problem of K-Means is resolved and thus the method is suitable for large data sets.

In one method presented in [22], a B^+-tree-based structure is used to index the document feature vectors. For the purpose of clustering, each leaf node is a document and each nonleaf node n is the representative of the cluster containing the documents in the leaf nodes of the subtree rooted at n. Suppose a new document is inserted into the clustering tree, Similar to the CF-tree and the B^+-tree based method, this is a balanced tree, and has the problem of possibly forcing non-similar items into the same node or cluster.

In [23], the Suffix Tree Clustering (STC) algorithm was proposed. This algorithm does not treat a document as a set of words but rather as a string, making use of proximity information between words. STC relies on a suffix tree to identify sets of documents that share common phrases and uses this information to create clusters and to summarize their contents for users. However, the time complexity of STC is quite high, which is a limitation for incremental or online Web document clustering. In their prototype clustering Web search engine, the document set size is restricted to a limit of 300 documents.

We propose a tree structure called a DC-tree which has some similarity to the CF-tree. The DC-tree node also stores a summary information of the cluster and a node entry can absorb an incoming entry when they are regarded as similar. However, the DC-tree aims at Web-page classification and hence the summary information is different from that of the CF-tree. Another major difference is that the DC-tree may not be a height balanced tree. Since the DC-tree is proposed for document clustering, the cluster representation, the clustering quality measurement and cluster-discovering algorithm are also different from BIRCH.

3 Feature Extraction and Document Clustering

Our first task is to identify a good feature extracting method suitable for the Web environment. In this section, we propose such a method. The document and document cluster representation will also be described. We describe three alternatives for measuring the similarity between two document feature vectors. Finally, the clustering quality evaluation method will be presented.

3.1 Statistical Analysis in the Web Domain

In order to study whether the traditional term weighting is suitable for the Web document clustering, we have carried out an experiment to analyze the number of distinct words and the average word frequency of the Web pages. We have used the stratified random sampling technique for collecting the document set in this experiment. Ten topics (Arts & Humanities, Business & Economy, Computers & Internet, Entertainment, Government, Health, Recreation & Sports, Science, Social Science, and Society & Culture) have been chosen from the *Yahoo!* search engine. Then we randomly collect 2000 documents for each topic. While collecting the

Web documents, the number of *valid words* [3] of the pages is counted. We will discard any document that contains zero valid word. Totally 20000 Web pages have been collected from the *Yahoo!* search engine. Some statistics about the words of Web documents are shown in Table 1.

	No. of distinct words	Average word frequency
Mean	174.75	1.68
Median	116.00	1.55
Standard Deviation	225.82	0.71
Maximum	4351	17.59
Minimum	1	1.0

Table 1: Statistics about the words of Web documents

We made a study on the number of distinct words in the Web documents. We notice that most of the Web documents contain less than 500 words. According to statistics, there are 17027 (85.14%) documents that contain less than or equal to 300 distinct words, 18930 (94.65%) documents that contain less than or equal to 500 distinct words, and 19712 (98.56%) documents that contain less than or equal to 1000 distinct words

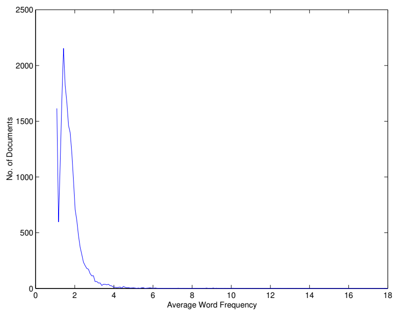


Figure 1: Document distribution along the average word frequency

According to Figure 1 and Table 1, the average word frequency (or term frequency) of almost all documents is less than 2.0 which means that most of words in a Web document will rarely appear more than 2 times. Therefore, the traditional term frequency weighting scheme which was designed for longer documents may not be useful in Web Domain.

3.2 Document Feature Extraction

Luhn [13] proposed that the frequency of word occurrences in an article furnishes a useful measurement of word significance. Let us order the words by their frequency of occurrences, resulting in the *rank order*. According to *Zipf's Law* [25] (see also [5]), the product of the frequency of use of words and the rank order is approximately constant.

Traditionally, a document D is represented by a feature vector of the form $(d_1, \ldots, d_n)$ where d_i is the numeric weight for the i-th feature and n is the total number of features. Each feature corresponds to a word appearing in the training corpus after stemming the words and the

[3] Here, the *valid word* is define as the remaining words of the page after stop-word removal and word stemming process.

removal of the stop-words. The following is a common feature extracting method [15] which is based on the number of occurrences of particular terms in the documents: (1) Eliminate common function words from the document texts by consulting a special dictionary, or stop-list, containing a list of high-frequency function words. (2) Compute the **term frequency** tf_{ij} for all remaining terms T_j in each document D_i, specifying the number of occurrences of T_j in D_i. Another method combines the term frequency with the inverse document frequency (TF-IDF). The **document frequency** df_j is the number of documents in a collection of N documents in which term T_j occurs. A typical inverse document frequency (*idf*) factor [16] of this type is given by $log(N/df_j)$. The weight of a term T_j in a document D_i is given by $w_{ij} = tf_{ij} \times log(\frac{N}{df_j})$. Then n terms with the highest weights in all the documents are chosen as the n features.

However, from experiments (see Section 3.1), we find that the above feature extracting methods are not suitable to the Web document clustering algorithm due to the following reasons: (1) As shown in Section 3.1, a Web document usually contains a small number of words (about 100-300 words). So, the term frequency of the key word may not be higher than that of the non-key word. (2) Usually, the users do not want too many clusters in the clustering result. However, the TF-IDF method may extract those features with relatively low document frequency because the lower document frequency will give a higher weight of the feature. Such features can lead to a clustering result containing many small clusters. (3) The above methods also ignore the *coverage* of the features extracted. The **coverage** of the features is defined as the percentage of documents containing at least one feature of the features extracted. Since the above methods only choose the best n terms, there is no guarantee that the n terms cover a high percentage of the documents. If the *coverage* is too low, there will be many documents represented by a feature vector with zero weight in all entries.

We propose a feature extracting method for the Web document clustering algorithm which does not depend on the term frequency. This method balances the trade-off between the coverage and the number of features used for document representation. In our problem domain the clustering aims to aid in information retrieval in Web searching by narrowing down the search scope. In such a scenario, the user may not want too many clusters in the result. Also, very large clusters or very small clusters are not desirable. Very large clusters cannot help to narrow the search scope. Very small clusters can increase the total number of clusters, and it may actually be caused by noise. The parameter k is used to set an *approximate* number on the cluster size. Hence the number of clusters is approximately N/k, where N is the total number of documents. The proposed method involves the following steps:

1. Randomly select a subset of documents with size m from the corpus.
2. Extract the set of words that appear at least once in the documents. Remove stop words and combine the words with the same root by using the stemming technique.
3. Count the document frequency of the words which are extracted in Step 2.
4. Set $lower = k$ and $upper = k$
5. Select all words with document frequency in the range from $lower$ to $upper$

6. Check if the *coverage* of these words is larger the pre-defined threshold. If so, stop. If no, set $lower = lower - 1$ and $upper = upper + 1$, and then goto Step 5.

Since the document set is usually very large, it is inefficient to carry out the feature extraction process on the entire document set. In order to extract the representative features from the documents, we randomly select a set of sample documents for feature extraction in Step 1. We have carried out some experiments (see Section 5.3) which show that this feature extraction method can extract a set of good features for Web document clustering. A stop-word list is used to remove the meaningless words which are not suitable to represent the concept of a class. Stemming technique is used to combine those words in similar form.

Since shorter feature vectors lead to shorter clustering time, Steps 4 to 6 try to minimize the number of features and obtain reasonable coverage for the features. Assume the user wants the resulting cluster to contain about k documents. In the ideal case, a feature for a cluster will appear only in the cluster and hence the document frequency of the feature is k. Therefore, we first select the features with document frequency equal to k, by setting *lower* and *upper* to k in Step 4. The range { *lower, upper* } is enlarged repeatedly in Step 6 to ensure sufficient coverage for the resulting feature set. From the above, we see that N/k is only a rough guideline, the actual number of clusters of the clustering result may not be the same as N/k. The method also makes use of a **coverage threshold** to ensure that the features selected has sufficient coverage. In our experiments (see Section 5), we find that 0.8 is a good coverage threshold value.

3.3 Document Representation

In our algorithm, a document (D_i) is represented in the following form: $D_i = (W_i, ID_i)$, where ID_i is the document identifier which can be used to retrieve document (D_i) and W_i is the feature vector of the document: $W_i = (w_{i1}, w_{i2}, \ldots, w_{in})$. Here n is the number of extracted features, and w_{ij} is the weight of the j-th feature, where $j \in \{1, 2, ..., n\}$. In our algorithm, a binary weighting scheme has been used. That is, w_{ij} is equal to 1 if D_i contains the j-th feature. Otherwise, w_{ij} is equal to 0. As mentioned in Section 3.2, a Web page typically does not contain many words so that the term frequency of a word cannot indicate the actual importance of this word. Therefore, the binary weighting scheme is more suitable to our problem domain.

3.4 Document Cluster (DC)

A **Document Cluster** value (DC) is a triple storing the information that we maintain about a set of documents within the same cluster: (1) the number of documents, (2) the set of document identifiers, and (3) the feature vector of the cluster.

Definition 1 *(DC Definition) Given N documents in a cluster:* $\{D_i, D_2, \ldots, D_N\}$*, the Document Cluster (DC) entry of a node is defined as a triple: DC = (N,ID,W), where N is the number of documents in the cluster, ID is the set of the document identifiers of the documents in the cluster, i.e.* $ID = \{ID_1, ID_2, \ldots, D_N\}$*, and W is the feature vector of the document cluster, i.e.* $W = (w_1, w_2, \ldots, w_n)$*, where* $w_j = \sum_{i=1}^{N} w_{ij}$*, and n is the number of extracted features.*

This triple not only summarizes the document frequency inside the cluster, but also can be used to measure the closeness between two clusters. The closeness or similarity measurement between two clusters will be described in Section 3.5.

The following lemma gives us the flexibility to combine two clusters into one and gives a DC value for the combined cluster.

Lemma 1 *(DC Additivity) Let* $DC_1 = (N_1, ID_1, W_1)$*, and* $DC_2 = (N_2, ID_2, W_2)$ *be the DC entries of two disjoint document clusters, where disjoint means a document does not belong to more than one cluster at the same time. Then the new DC entry,* DC_{new}*, of the cluster that is formed by merging the two disjoint clusters, is:* $DC_{new} = (N_1+N_2, ID_1 \cup ID_2, W_1+W_2)$*, where* $W_1 + W_2 = (w_{11} + w_{21}, w_{12} + w_{22}, \ldots, w_{1n} + w_{2n})$*, and* n *is the number of extracted features.*

3.5 Similarity Model

We will consider three alternatives for measuring the similarity between two document feature vectors based on the classical vector space model [17]. The first measure is called *cosine correlation* which is commonly used to measure the similarity between the feature vectors of two documents (e.g. [9, 10, 12, 21]). A larger cosine value indicates that they are more similar. The second similarity measuring method uses the *Euclidean distance*, which is the most common similarity measuring method in high dimensional data clustering [6, 14, 24]. The third alternative is the *Manhattan distance.*

Given two Document Clusters: $DC_1 = (N_1, D_1, W_1)$ and $DC_2 = (N_2, D_2, W_2)$, the definitions of the distance and/or similarity of the three methods are shown in the following table.

Alternatives	Distance(DC_1, DC_2)	Similarity(DC_1, DC_2)
Cosine correlation		$\frac{\sum_{i=1}^{n} \frac{w_{1i}}{N_1} \times \frac{w_{2i}}{N_2}}{\sqrt{\sum_{i=1}^{n} (\frac{w_{1i}}{N_1})^2} \sqrt{\sum_{i=1}^{n} (\frac{w_{2i}}{N_2})^2}}$
Euclidean	$\sqrt{\sum_{i=1}^{n} (\frac{w_{1i}}{N_1} - \frac{w_{2i}}{N_2})^2}$	$1 - \frac{1}{n}\sqrt{\sum_{i=1}^{n} (\frac{w_{1i}}{N_1} - \frac{w_{2i}}{N_2})^2}$
Manhattan	$\sum_{i=1}^{n} \lvert\frac{w_{1i}}{N_1} - \frac{w_{2i}}{N_2}\rvert$	$1 - \frac{\sum_{i=1}^{n} \lvert\frac{w_{1i}}{N_1} - \frac{w_{2i}}{N_2}\rvert}{n}$

In the above, n is the number of extracted features; $W_1 = (w_{11}, w_{12}, \ldots, w_{1n})$; $W_2 = (w_{21}, w_{22}, \ldots, w_{2n})$. Note that the Manhattan distance requires much simpler mathematical operations compared to the other two distances.

3.6 Evaluation Techniques

In order to evaluate the quality of the clustering result, we adopt the *F-Measure* evaluation technique which was first introduced in [19] (see also [11]). The details of this evaluation methodology is described below.

For each hand-labeled topic T in the document set, suppose that a cluster X corresponding to that topic is formed.

N_1 = no. of documents of topic T in cluster X

N_2 = no. of documents in cluster X

N_3 = total no. of documents of topic T

$P = \text{Precision}(X,T) = \frac{N_1}{N_2}$

$R = \text{Recall}(X,T) = \frac{N_1}{N_3}$

The F-measure [11, 19] for the topic T is defined as below:

$$F(T) = \frac{2PR}{P+R}$$

With respect to a topic T, we consider the cluster with the highest F-Measure to be the cluster C for T, and that F-Measure becomes the score for topic T. The overall F-Measure [10] for the tree clustering result is the weighted average of the F-Measure for each topic T:

$$Overall\_F\_Measure = \frac{\sum_{T\in M}(|T| \times F(T))}{\sum_{T\in M}|T|}$$

where M is the set of topics, $|T|$ is the number of documents of topic T, and $F(T)$ is the F-Measure for topic T.

4 DC-tree

In this section we propose a Web document clustering algorithm by means of a Document Cluster tree (DC-tree). In the DC-tree, each node can be considered as a document cluster. The tree structure is used to guide the incoming document object to an appropriate Document Cluster (DC) at the leaf nodes. It is similar to the B^+-tree [22] in that the index records in its leaf nodes contain pointers to data objects, but it is not a height-balanced tree. This structure is designed so that assigning a document into a cluster requires visiting only a small number of nodes.

A **DC-tree** is a tree with four parameters: branching factor (B), two similarity thresholds (S_1 and S_2, where $0 \leq S_1, S_2 \leq 1$) and the minimum number of children of a node (M). A non-leaf node contains at most B entries of the form $(DC_i, Child_i)$, where $i = 1, 2, \ldots, B$, "$Child_i$" is a pointer to its i-th child node or a document, and DC_i is the DC entry of the sub-cluster represented by its i-th child or a document. So, a non-leaf node represents a cluster that is made up of all the sub-clusters represented by its entries. A DC leaf node contains at most B entries, each is an entry of the form (DC_i, Doc_i), where $i \in \{1, 2, \ldots, B\}$, "$Doc_i$" is a pointer to a document or a set of documents, and DC_i is the DC entry of the corresponding sub-cluster. We call the set of documents under such a pointer a **document leaf node**, to distinguish it from the tree leaf node or **DC leaf node** (see Figure 2). A DC leaf node also represents a cluster made up of all the sub-clusters represented by its DC entries. The DC-tree allows an incoming document entry to be inserted in a new document leaf node at different levels of the tree. So, the DC-tree is not a height balanced tree. Figure 2 shows a sample DC-tree with a height of 2, $B = 3$, $M = 2$. Note that the tree is not balanced. In the tree construction, two thresholds are used:

1. **Threshold** S_1: In order to prevent poor document clustering result (i.e. documents in different classes are assigned to the same subtree or cluster) caused by document insertion

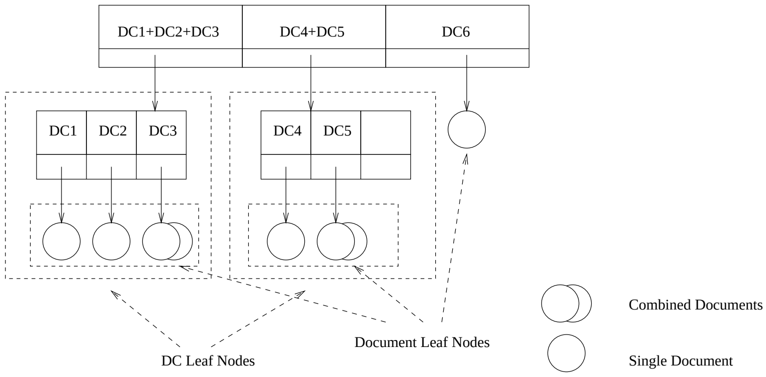


Figure 2: A DC-tree Example

order, S_1 is used to determine whether the incoming document entry E can be passed into the next level during the insertion process. If there exists an child entry of the current node such that the similarity value between this entry and the incoming entry is higher than the S_1, the incoming entry can be passed to the corresponding child node. Otherwise, the incoming entry will be added to the current node as a new document leaf node.

2. **Threshold** S_2: Since the DC-tree is used for document clustering but not indexing, it is not necessary to force each leaf entry to point to a single document. In order to reduce the insertion time, the incoming document entry can be **combined** with a leaf entry, if their similarity value is higher than S_2. This entry combination makes use of the merging as described in Lemma 1, it can save on node insertion and splitting operations so that the insertion time can be reduced.

The DC-tree is a compact representation of the dataset because each entry in a leaf node is not a single data point but a cluster of data points (the leaf node absorbs as many data points as the threshold S_2 allows). With the above DC-tree definition, the tree size will be a function of the threshold values S_1 and S_2. In general, the tree height increases with S_1 and the tree size increase with S_2. If we set S_1 to zero and S_2 to any number larger than one, the DC-tree will be similar to a balanced tree like a B^+-tree or an R-tree.

4.1 Insertion

We now present the algorithm for inserting a document object into a DC-tree. Here the document object can be either a single document, or a cluster of documents which is represented by a DC entry (E). If the document object is a single document, it will be first encapsulated into a DC entry (E) (i.e E contains only one document). The insertion algorithm proceeds in the steps as shown below:

1. **Identifying the appropriate leaf node:** Starting from the root, E recursively descends the DC-tree by choosing the *closest* child node with the similarity value higher than S_1. If such child node does not exist, E is inserted as a new document leaf node to an empty entry of the node. If there is no such empty entry, node *splitting* is required (see Section 4.2).

2. **Modifying the leaf node:** When we reach a DC leaf node, we find the *closest* leaf entry to E, say L_i, and then tests whether it can be combined with E without violating the similarity threshold requirement of S_2. If so, the DC entry for L_i is updated to reflect this combination. Note that the DC entry of the new cluster can be computed from the DC entries for L_i and E according to Lemma 1. Otherwise, E is added to the leaf node. If there is space in the leaf node for this new entry to fit in, we are done, otherwise we must *split* the leaf node. Section 4.2 will discuss the node splitting algorithm in detail.

3. **Modifying the path from the leaf node to the root:** After E is inserted into a leaf node, we must update each non-leaf entry on the path from the root leading to the leaf node. In the absence of a split, this simply involves adding DC entries to reflect the addition of E. A leaf node split requires us to insert a new non-leaf entry into the parent node, which points to the newly created leaf node. If the parent node has space for this entry to fit in, then at all higher levels, we only need to update the DC entries to reflect the addition of the E. In general, however, we may have to *split* the parent node as well, and so on up to the root. If the root is split, the height of the tree will be increased by one and a new root will be created.

4.2 Node Splitting

In order to add a new entry to a full node containing B entries, it is necessary to divide the collection of the $B+1$ entries between two nodes. The division should be done in a way such that the similarity between the two new nodes will be minimized and the similarity among the documents within the same node will be maximized. We now turn to algorithms for partitioning a set of $B+1$ entries into two groups, one for each new node. The most straightforward way to find the optimal node split is to generate all possible groupings and choose the best. However, the number of possibilities is very large (i.e. approximately 2^{B-1}). Below is the node splitting algorithm used in our DC-tree insertion algorithm. This node splitting algorithm is similar to a method considered in R-tree [7].

1. **Pick a seed for each group:** For each pair of entries E_1 and E_2, compute the similarity between them. Choose the pair with the lowest similarity value as the first elements of the groups. Resolve ties by choosing the pair with the largest number of documents.

2. **Check if done:** If all entries have been assigned, stop. If one group has so few entries that all the rest must be assigned to it in order for it to have the minimum number M, assign them and stop.

3. **Select entry to assign:** For each entry E not yet in a group, calculate the similarity between E and the seed entry of each group. Assign the entry with the highest similarity value into the corresponding group. Goto step 2.

This splitting algorithm attempts to find a good node split in quadratic time in terms of B, but it is not guaranteed to find the optimal solution.

4.3 Deletion and Node Merging

The data deletion algorithm is similar to that of the B^+-tree. If the number of remaining entries is larger than or equal to the minimum entries number parameter M after removing an entry, the deletion operation is done. Otherwise, node-merge operation should be carried out. That is, the underflow node will be merged with its sibling. Further node-merge operation is needed when the entry underflow phenomenon occurs in the parent node. This operation can be propagated to the root node where the height of tree may be reduced if necessary.

4.4 Identifying the Interesting Clusters

The cluster identifying process starts from the root of the tree. A breath-first-search algorithm is applied to discover the interesting clusters. An interesting cluster is defined as a cluster that contains representative features and whose size is in a predetermined range. We can make use of the *lower* and *upper* values found in our proposed feature extracting method (see Section 3.2) to determine the cluster size range. Let l and u be the lower bound and upper bound of the cluster size range, then l and u can be determined by the following equations: (1) $l = lower \times \frac{N}{m}$, (2) $u = upper \times \frac{N}{m}$, where N is the data set size and m is the size of the sample data set used in the feature extraction. Also, the range can be adjusted manually to obtain a good clustering result. Once we have identified an interesting cluster, the sub-clusters in its descendent nodes will not be scanned. A **representative feature** is defined as a feature that has enough support in the cluster. That is, the document frequency of the representative features must be greater than some user-defined threshold. We call this threshold the **representative threshold**. These features will then be used to represent the cluster.

5 Experimental Results

In this section, we list some experimental results for the proposed feature extracting method and the Web document clustering algorithm. We have applied the algorithm on both synthetic data and real data for studying the performance and the relationship among the parameters of the proposed algorithms. We find that the behaviour of the algorihtms are similar for the real data and synthetic data. For interest of space, we shall report on either data set in each performance analysis. All experiments are run on the *Sun Ultra 5/270* with 512 MB of RAM. The programs are compiled using *GNU* C++ and executed on the *Solaris 2.6* operating system.

5.1 Synthetic Data Generation

The synthetic datasets are generated by a generator that we have developed. The data generation is controlled by four parameters: (1) the number of document feature vectors, (2) the number of features, (3) the number of clusters, and (4) the **similarity level** L which is ranged from 0 to 1. Given x feature vectors and y clusters, for each cluster C, we set x/y features F randomly to represent the cluster. For each document generated for C, the features in F will have a chance of L being set to 1 and $1 - L$ being set to 0. The features not in F will have a chance of $1 - L$ being set to 1. Hence the higher the similarity level, the more similar the documents will be within the same cluster.

5.2 Yahoo! Web Page Collections

We have collected two sets of Web pages from the *Yahoo!* search engine [20]. One contains ten non-correlated topics while another contains ten correlated topics. The Web pages indexed by *Yahoo!* search engine have already been classified.

To construct the non-correlated document set, we have selected ten sub-topics, namely, author, credit bank, internet game, wine, film, disease organization, job opportunity, soccer, astronomy organization, psychology department. For each topics, we collected 2000 Web pages. So, there are totally 20000 Web pages in this dataset.

For the correlated topic Web document set, we have selected ten sub-topics from the following location:

`http://dir.yahoo.com/Computers_and_Internet/Programming_Languages/`

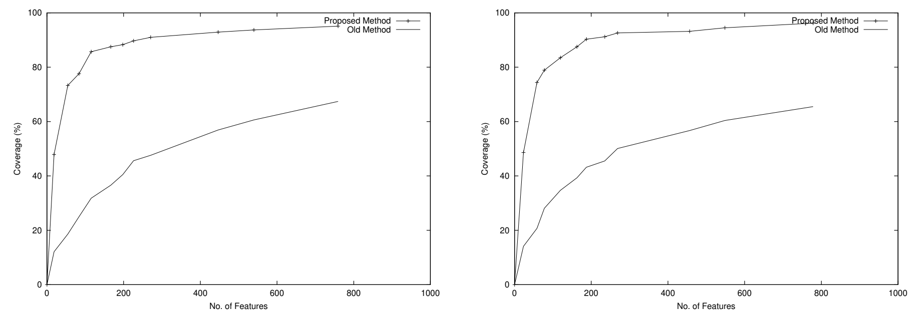
from *Yahoo!* search engine, namely, COBOL, Fortran, Java, JavaSript, Lisp, Pascal, Perl, Python, Smalltalk and Visual Basic. For each topics, we collected 2000 Web pages. So, there are also totally 20000 Web pages in this dataset.

5.3 Performance of the Feature Extraction method

As mentioned in Section 3.2, traditional extracting method ignores the coverage of the features extracted. In order to compare our proposed feature extracting method (or proposed method) with the traditional method (or old method) in terms of the coverage, we carry out an experiment to record the coverage value for different number of features extracted by these two methods.

For each real data set (see Section 5.2), we have collected 5 samples with size 1000 documents. In order to create a sample with size 1000, we randomly select 100 pages from each sub-topic document set. For the proposed method, we set parameter k as 20, that is, the word with document frequency near to 50 will be extracted. Then, we record the coverage value and the number of features n found by varying the range $\{lower, upper\}$. For the old method, we first calculate the weights of all features of the documents by using the TF-IDF method (see Section 3.2). Since a feature can appear in more than one document, it will have more than one value of the weight. In this case, we choose the largest weight as the weight of this feature. We then sort the features in descending order according to their weights. We choose the first n features from the ordered list and calculated the coverage of these n features. Figure 3 (a) shows the plot of the average coverage against the average number features for the non-correlated topic documents set. Figure $3(b)$ shows the plot of the average coverage against the average number features for the correlated topic documents set.

From both Figure $3(a)$ and (b), we can notice that the proposed feature extracting method always gives a larger coverage than the old method when the same number of features are used.



(a) (b)

Figure 3: Coverage against number of features for (a) non-correlated topic and (b) correlated topic document sets

5.3.1 Clustering Quality

In order to study the influence of the coverage factor on the clustering quality, we use the fourth feature set (i.e., the features extracted when $lower = 47$ and $upper = 53$) for clustering the two document sets mentioned in Section 5.2. Table 2 and Table 3 show the number of features and the corresponding coverage of the feature set for the five samples of the non-correlated topic and the correlated topic document sets respectively.

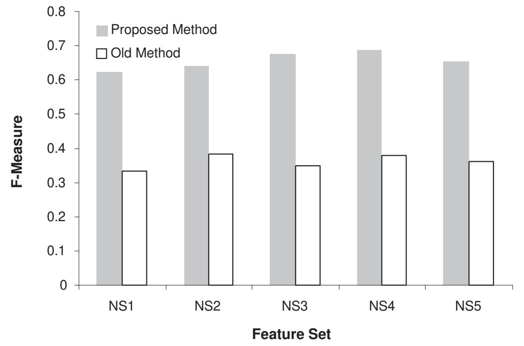
	No. of Features	Proposed Method Coverage (%)	Old Method Coverage (%)
Sample 1 (NS1)	113	83.90	31.20
Sample 2 (NS2)	116	85.40	31.70
Sample 3 (NS3)	117	86.90	32.30
Sample 4 (NS4)	119	89.10	33.10
Sample 5 (NS5)	112	83.20	30.90
Average	115.30	85.70	31.80

Table 2: Feature extracting result of the five samples for the non-correlated topic document set

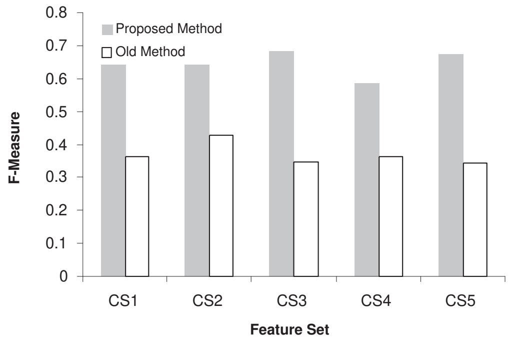
	No. of Features	Proposed Method Coverage (%)	Old Method Coverage (%)
Sample 1 (CS1)	117	81.70	34.00
Sample 2 (CS2)	117	82.10	33.10
Sample 3 (CS3)	127	87.70	36.90
Sample 4 (CS4)	114	79.20	33.10
Sample 5 (CS5)	123	88.44	35.70
Average	119.60	83.50	34.70

Table 3: Feature extracting result of the five samples for the correlated topic document set

We apply our proposed document clustering algorithm with these features to clustering the above mentioned two real data sets. The parameters of the DC-tree used in the clustering algorithm are: minimum number of entries = 8, branching factor = 20, similarity threshold S_1 = 0.3, similarity threshold S_2 = 0.8. These parameters have been chosen by trial and error, meaning that we have tried different combinations of values and discovered that the chosen values gives comparably good results. In this experiment, we use the binary weighting scheme to construct the document feature vectors. We evaluate the clustering quality in terms of F-Measure. Figure 4(a) and (b) show the experimental results for the non-correlated topic and correlated topic document sets respectively.



(a)



(b)

Figure 4: F-Measure for (a) non-correlated topic and (b) correlated topic document set

According to Figure 4(a) and (b), we can notice that the features found by the proposed feature extracting method always gives a better document clustering result than those extracted by old method. It is because the coverage of the feature set extracted by the proposed method is always larger than that of feature set extracted by the old method. Low coverage will cause many of document vectors to contain all zero entries. Such document vectors will always be assigned to the same cluster even though they come from different classes. So, low coverage can lead to poor clustering result.

5.3.2 Binary Weight vs Numerical Weight

As shown in Section 3.1, most of key words ($> 90\%$) in a Web document appear less than 2 times. In this experiment, we want to make a comparison between the performance of the document clustering in Web domain using the binary weighting scheme and the numerical weighting scheme. We use the same five feature sets and data sets of last experiment. The DC-tree parameters are the same as last experiment. We employ both binary and numerical weighting schemes to construct the document vectors and apply our proposed document clustering algorithm to cluster these document vectors. We compare their performance in terms of clustering quality (i.e. the F-Measure value of the clustering result). The experimental results on the non-correlated topic document set and the correlated topic document set are shown in Figure 5(a) and (b) respectively. Both figures show that binary weighting scheme always gives a better clustering result than the numerical one does.

5.4 Alternative Similarity Measurements

In Section 3.5, we have defined three methods for measuring the similarity between the feature vectors of two documents, namely the *cosine correlation*, the *Manhattan distance* and the

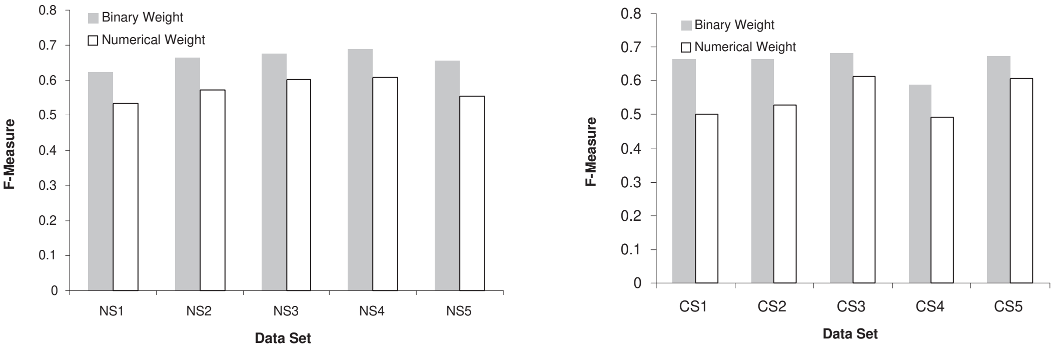
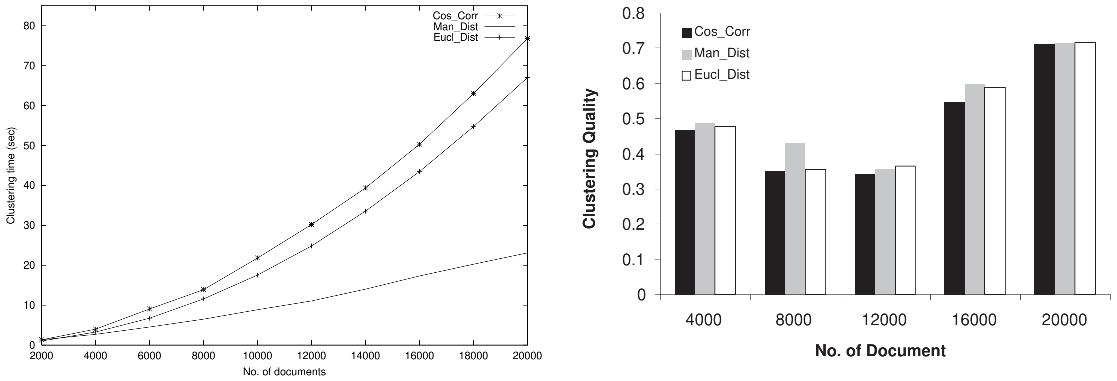


Figure 5: Comparison between the binary weighting scheme and the numerical weighting scheme in terms of F-Measure for (a) non-correlated topic and (b) correlated topic document set

Euclidean distance. In order to investigate the influence of these three similarity measurement functions on the performance of the algorithm, the clustering time of the program using different similarity function is plotted against the number of documents which is shown in Figure 6(a). This experiment is carried out on a set of synthetic data of 20000 web pages as mentioned above. We have used 100 features, ten clusters and a similarity level of 0.8. The parameters of the DC-tree used in the clustering algorithm are: minimum number of entries = 8, branching factor = 20, similarity threshold $S_1 = 0.3$, and similarity threshold $S_2 = 0.8$. These parameters have been chosen by trial and error, meaning that we have tried different combinations of values and discovered that the chosen values gives comparably good results.



(a) (b)

Figure 6: Comparison among three similarity functions

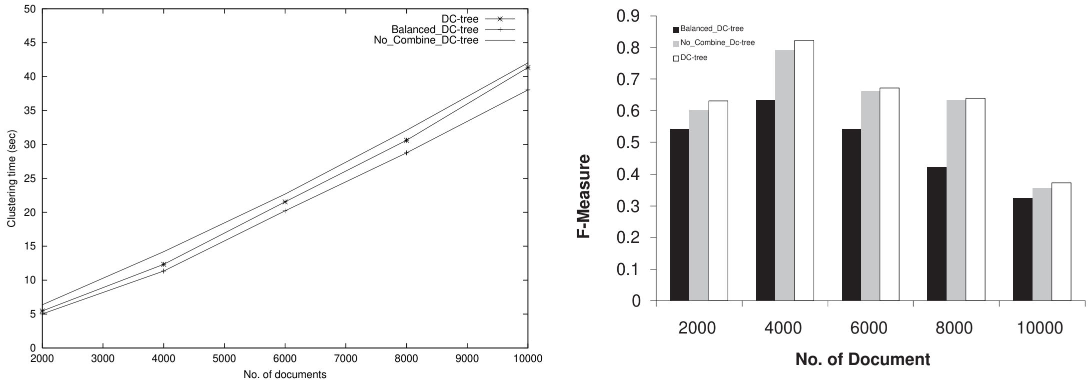
From Figure 6(a), the *Manhattan distance* method always gives the best performance in terms of clustering time since its similarity measurement computation is simple. Figure 6(b) shows that the three similarity functions give similar clustering results in terms of the F-Measure value. Therefore, we choose the *Manhattan distance* method as the similarity evaluation method in our remaining experiments.

5.5 DC-tree Characteristics

For the DC-tree, each incoming entry is not forced to be inserted into a leaf node in DC-tree and this should make the clustering result less sensitive to the insertion order. Moreover, the leaf entry can include an incoming entry when they are regarded as very similar to each other to reduce node splitting as well as maintaining good clusters. In order to measure the impact of these features of the DC-tree on the document clustering effectiveness and efficiency, we carry out two experiments in which two variants of the DC-tree are created.

The first variant is called the *Balanced_DC-tree*. With this variant, the tree is balanced and an incoming entry will always be inserted into the leaf node, so the threshold S_1 is not used. Like the DC-tree, an incoming entry will be combined with a leaf entry if their similarity value is larger than the threshold S_2. For the second variant called *No_Combine_DC-tree*, the tree is not balanced but each leaf entry corresponds to only one Web page or document, so an incoming entry will not be combined with a leaf entry even if they are very similar to each other. Hence S_1 is used but S_2 is not used. More node-splitting would be expected in this case.

We compare the effectiveness among the document clustering algorithm using DC-tree, using *Balanced_DC-tree* and using *No_Combine_DC-tree* in terms of the F-Measure value of the clustering result. Synthetic data is used in this experiment. We also compare the efficiency in terms of the clustering time. The result, as shown in Figure 7 (a) and (b), shows that DC-tree always gives better performance and clustering results.[4]



(a) (b)

Figure 7: Comparison between DC-tree and its variants

5.6 Comparing DC-tree and B^+-tree

In this experiment, we want to make a comparison between the effectiveness of the document clustering in Web domain using the DC-tree and the B^+-tree. We use the two real data sets of 20000 web documents each for both the correlated topics and non-correlated topics. We use two of the feature sets (i.e., Sample 4 in Table 2 and Sample 4 in Table 3) found in the experiment 5.3 for converting each document to a feature vector in this experiment. This set

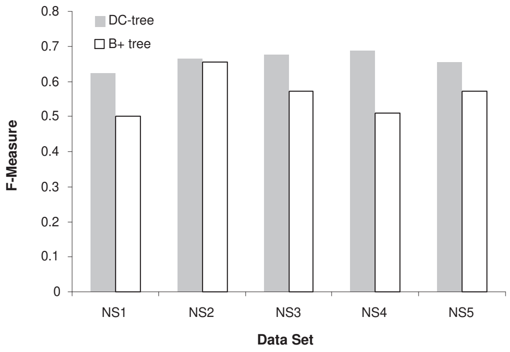
[4]Note that the results are different from that shown in Figure 6 because a different set of synthetic data is used. However the range of the F-measure values is about the same.

contains 119 features. The DC-tree parameters are the same as the previous experiment. We employ the binary weighting schemes to construct the document vectors and apply the document clustering algorithms to cluster these document vectors. We compare their effectiveness in terms of the F-Measure value of the clustering result and also on a precision value.

	non-correlated topics	correlated topics
No. of clusters found	10	10
Cluster size range	1000-3000	1000-3000
Representative threshold	0.45	0.4
Clustering time	94.21 seconds	72.27 seconds
F-Measure Value	0.6753	0.6453

Table 4: Real data clustering results

The representative thresholds for the DC-tree method in the two cases are shown in Table 4, and are chosen by trial and error. The table also shows some experimental measurements. The experimental results in terms of F-measure on the non-correlated topic document set and the correlated topic document set are shown in Figure 8(a) and (b) respectively. The figures show that DC-tree always gives a better clustering result than the B^+-tree does for both real data sets in terms of F-measure.



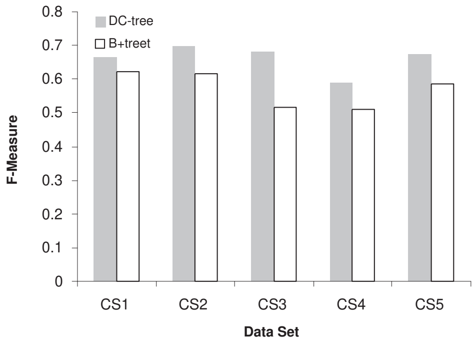


Figure 8: F-Measure for (a) non-correlated topic and (b) correlated topic document set

5.6.1 Precision

For each cluster found, we calculate the document support of each known topic, which is the number of documents containing the same topic. Then we choose the topic with the highest document support as the topic of the cluster. Tables 5 (a) and (b) show the precision values of the clusters found by these two algorithms. They show that the DC-tree always gives a better clustering result than B^+-tree in terms of the precision of the resulting clusters. The results are also quite high compared to results reported in [23] in terms of the average precision. In [23], the precision of the best precision results shown are about 0.4. Though we cannot compare directly with [23] since the data sets used are different, the results are very encouraging.

Table 6 and Table 7 also lists the representative features of each cluster. Since the classification is automatic, we do not have a simple title for each topic of the clusters found, but we can guess the topic of the clusters by their representative features. The representative features (words) are used to indicate to the users what each cluster is about.

Cluster Topic	Precision	
	DC-tree	**B^+-tree**
author	0.61	0.59
internet game	0.73	0.55
wine	0.63	0.65
credit bank	0.66	0.57
soccer	0.73	0.47
astronomy organization	0.88	0.77
psychology department	0.85	0.63
job opportunity	0.91	0.72
film	0.72	0.73
disease organization	0.64	0.65
Average	**0.736**	**0.633**

(a)

Cluster Topic	Precision	
	DC-tree	**B^+-tree**
Cobol	0.64	0.54
Fortran	0.72	0.72
Java	0.73	0.62
JavaScript	0.64	0.67
Lisp	0.73	0.63
Pascal	0.81	0.67
Perl	0.78	0.72
Python	0.82	0.55
Smalltalk	0.61	0.57
VisualBasic	0.77	0.65
Average	**0.725**	**0.634**

(b)

Table 5: Precision of the clusters (a) non-correlated topics (b) correlated topic

Cluster Topic	No. of Documents	Features
author	2157	book, story, year
internet game	1988	best, free, game, review
wine	2461	wine
credit bank	2302	credit, service, union
soccer	1312	report, soccer, team
astronomy organization	1712	astronomy,union
psychology department	1892	psychology, university
job opportunity	1939	career, experience, job, location
film	1664	film, star
disease organization	2245	disease, health, medical, organization

Table 6: The clusters found for non-correlated data

5.7 Incremental updates on Real Data Set

In order to study the impact of the incremental updating the clustering result of the non-correlated topic document set in the last experiment, we insert 5000 additional documents into the existing DC-tree and observe the change of clustering quality. 433.22 seconds are used to convert the 5000 documents into feature vectors. Table 8 shows the changes of clustering quality for each 1000 additional documents added to the existing clustering result. From the table, we notice that the F-Measure value decreases from 0.6753 to 0.6010. The clustering quality only decreases about 11% in terms of F-Measure value when 25% more documents are added to the existing document set. Note also that each update in the document set will only involve a very small amount of computation in the clustering. Therefore, our document clustering algorithm is suitable for incremental updating in Web pages.

Cluster Topic	No. of documents	Features
Cobol	2232	cobol, coughlan, procedure
Fortran	2021	eacute, fortran
Java	2403	applet, class, event
JavaScript	2305	active, document, javascript
Lisp	1606	lisp, object
Pascal	1884	define, interface, pascal, xplain
Perl	1422	cpan, module, perl, port
Python	1810	kendall, python, thread
Smalltalk	1991	browse, develop, message, smalltalk
VisualBasic	1641	basic, click, item, visual

Table 7: Classification result for the correlated topic Web page collection

No. of documents	F-Measure value
21000	0.6555
22000	0.6475
23000	0.6249
24000	0.6062
25000	0.6010

Table 8: Incremental Update Results

6 Conclusion

In this paper, we introduce a tree structure called DC-tree for the incremental and hierarchical Web document clustering. We propose a feature extracting method for document clustering in the Web domain. This method takes into account the special characteristics of Web pages, and the special requirements of the problem domain. We show by experiment that the features of the DC-tree give good results in terms of accuracy and efficiency.

References

[1] R. Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for date mining applications. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 94–105, Seattle, Washington, USA, June 1998.

[2] F. Can and N. D. Drochak II. Incremental clustering for dynamic document databases. In *Proceedings of the 1990 Symposium on Applied Computing*, pages 61–67, 1990.

[3] Moses Charikar, Chandra Chekuri, Tomas Feder, and Rajeev Motwani. Incremental clustering and dynamic information retrieval. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on the Theory of Computing*, pages 626–635, El Paso, Texas, USA, May 1997.

[4] Chun-hung Cheng, Ada Wai-chee Fu, and Yi Zhang. Entropy-based subspace clustering for mining numerical data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 84–93, San Diego, CA, USA, August 1999.

[5] R. A. Fairthorne. Empirical hyperbolic distributions (bradford-zipf-mandelbrot) for biliometric description and prediction. *Journal of Documentation*, 25:319–343, 1969.

[6] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. CURE: An efficient clustering algorithm for large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 73–84, Seattle, Washington, USA, June 1998.

[7] A. Guttman. R-tree: A dynamic index structure for spatial searching. In *Proceedings of ACM SIGMOD*, pages 47–56, 1984.

[8] K. Koperski, J. Adhikary, and J. Han. Spatial data mining: Progress and challenges. In *Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, Canada, June 1996.

[9] Wai Lam and Chao Yang Ho. Using a generalized instance set for automatic text categorization. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 81–89, Melbourne, Australia, August 1998.

[10] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 16–22, San Diego, CA, USA, August 1999.

[11] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12, Dublin, Ireland, July 1994.

[12] David D. Lewis, Robert E. Schapire, James P. Callan, and Ron Papka. Training algorithms for linear text classifiers. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 298–306, Zurich, Switzerland, August 1996.

[13] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–165, 1958.

[14] Raymond T. Ng and Jiawei Han. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th VLDB Conference*, pages 144–155, Santiago de Chile, Chile, September 1994.

[15] Gerard Salton. *Automatic Text Processing. The Transformation, Analysis, and Retrieval of Information by Computer*, chapter 9, pages 275–312. Addison-Wesley Publishing Company, 1989.

[16] Gerard Salton and M. J. McGill. *Introduction to Modern Information Retrieval*, page 63. McGraw-Hill, New York, 1983.

[17] Gerard Salton and M. J. McGill. *Introduction to Modern Information Retrieval*, pages 120–123. McGraw-Hill, New York, 1983.

[18] Seán Slattery and Mark Craven. Combining statistical and relation methods for learning in hypertext domains. In *Proceedings of the 8th International Conference on Inductive Logic Programming*, Madison, Wisconsin, USA, July 1998.

[19] C. J. van Rijsbergen. *Information Retrieval*. Butterworth & Co (Publishers) Ltd, 1979.

[20] Yahoo Search Engine. http://www.yahoo.com.

[21] Yiming Yang. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 13–22, Dublin, Ireland, July 1994.

[22] Clement T. Yu and Weiyi Meng. *Principles of Database Query Processing for Advanced Applications*, chapter 10, pages 359–385. Morgan Kaufmann Publisher, Inc., 1998.

[23] Oren Zamir and Oren Etzioni. Web document clustering: A feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 46–54, Melbourne, Australia, August 1998.

[24] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: An efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 103–114, Montreal, Canada, June 1996.

[25] H. P. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, Massachusetts, 1949.